An insight into lexicographic practices in Europe

Results of the extended ELEXIS Survey on User Needs

Carole Tiberius
Dutch Language Institute

Jelena Kallas / Margit Langemets
Institute of the Estonian Language

Svetla Koeva
Bulgarian Language Institute

Iztok Kosem
Jožef Stefan Institute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results of the extended **ELEXIS** Survey on **User** Needs

= Needs of lexicographers

= Needs of lexicographers at the observer institutions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

ELEXIS

H2020 project which has created an infrastructure for lexicography to

(1) enable efficient access to high quality lexicographic data so that it can also be used by other fields including Natural Language Processing (NLP), artificial intelligence (AI) and digital humanities, and

(2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources.

To gain more insight into current lexicographic practices, workflows and the specific needs of lexicographers, a number of surveys have been conducted.

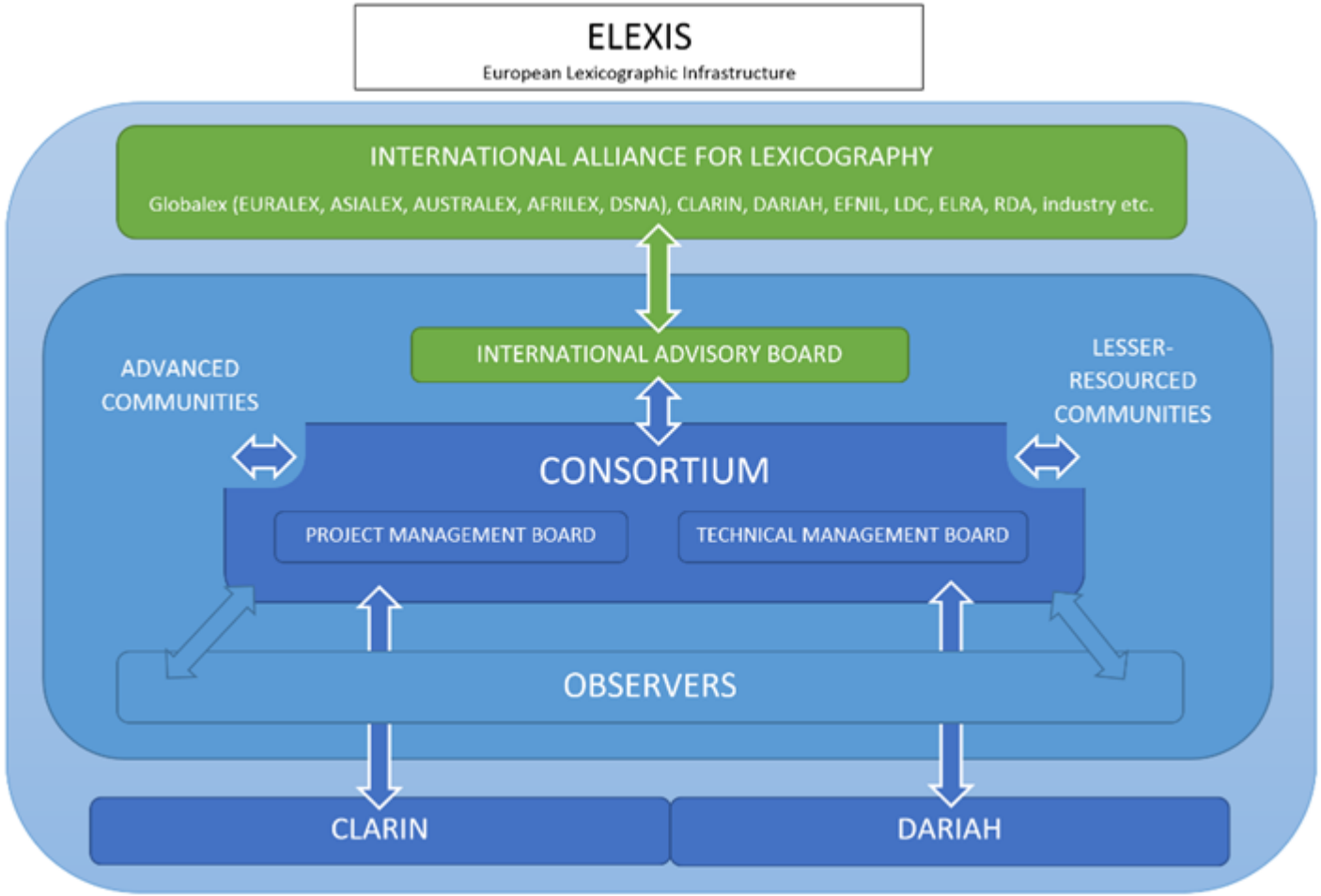
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS Surveys

- 2018 - Survey on Lexicographic practices: A Survey of Lexicographers' Needs (45 countries (36 European and 9 outside Europe))
- 2018 - Survey on Lexicographic practices: A Survey of Lexicographers' Needs for Lexicographic Partner Institutions (10 countries)
- 2020-2021 - Survey on Lexicographic practices: A Survey of Lexicographers' Needs for Observer Institutions (54 observers from 32 countries)

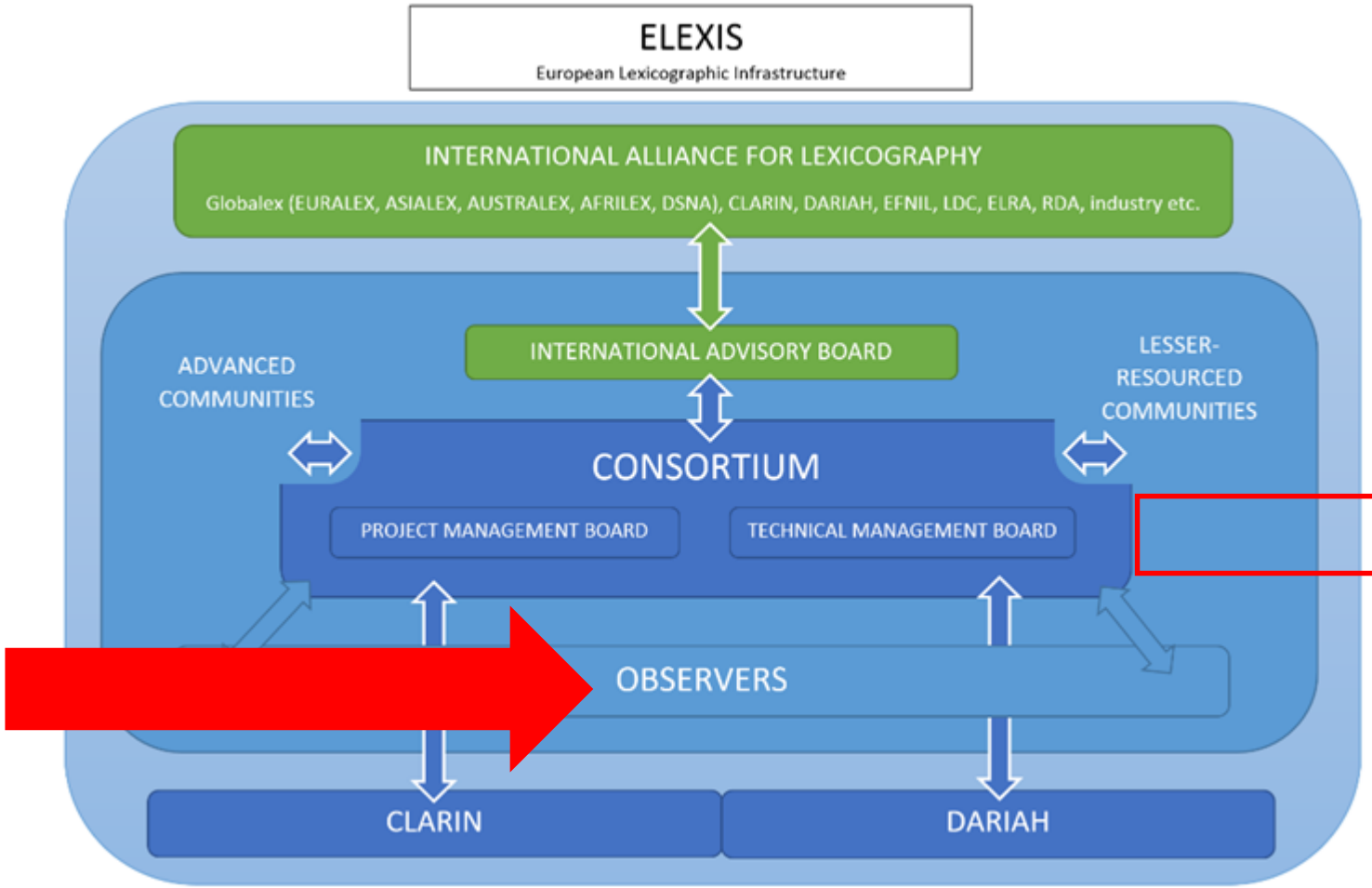
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS organisational structure



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS organisational structure



Lexicographic Partners

- Austrian Academy of Sciences
- Institute for Bulgarian Language Prof Lyubomir Andreychin
- Society for Danish Language and Literature
- Institute of the Estonian Language
- Trier University, Trier Center for Digital Humanities
- Hungarian Academy of Sciences, Research Institute for Linguistics
- K Dictionaries
- Dutch Language Institute
- Belgrade Center for Digital Humanities
- Jožef Stefan Institute
- Real Academia Española.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS observers

elexis european lexicographic infrastructure

elexis european lexicographic infrastructure

Home Resources Observers Events News About

OBSERVERS

The ELEXIS community is now made up of 17 partner and 56 observer institutions from 35 different countries.

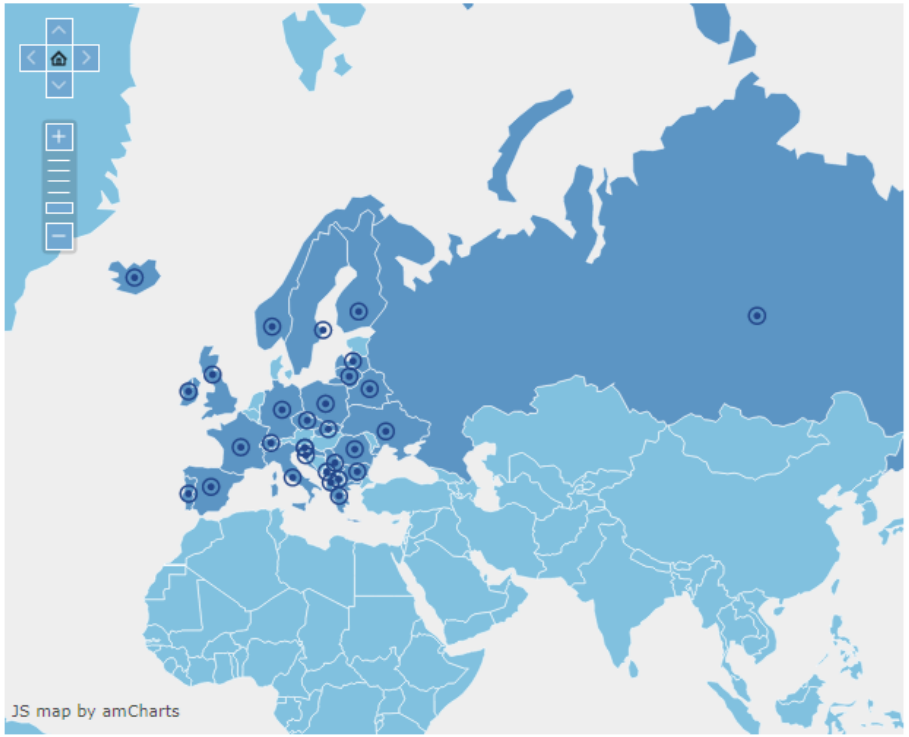
A list of current observer institutions can be found below.

If you wish to join the ELEXIS family as an observer to receive exclusive invitations and access to various services and tools, be informed of the most up-to-date developments in the field of lexicography, and several other benefits, you can still **apply for observer status**.

If you are looking for more in-depth information about the benefits that come with the observer status, please see our **benefits description**, **Vienna observer event recordings**, and **FAQ**.

OBSERVER STATUS

Learn about the benefits for your institution.



https://elex.is/observers/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

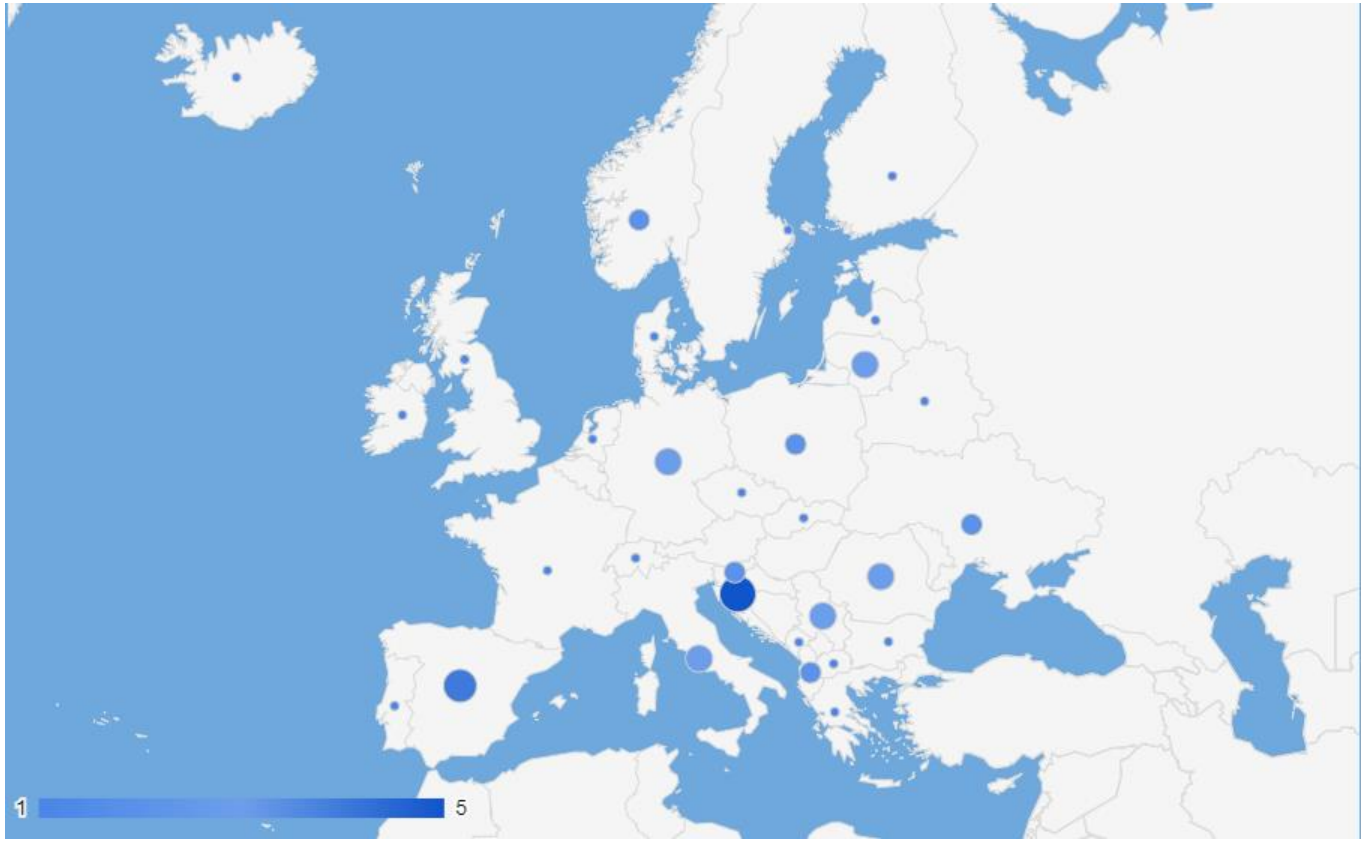
Survey for Observers

- Improved version of the survey for lexicographic partners
- 121 questions in 6 sections:
  - (1) General information
  - (2) Types of lexicographic resources. Software and tools supporting the workflow
  - (3) Publication and access. Crowdsourcing and Gamification
  - (4) Retrodigitised dictionaries
  - (5) Data formats. Metadata. Availability
  - (6) Past and Future.
- "yes/no" questions, multiple choice questions, open-ended questions. Not all questions were obligatory.
- One survey per institution
- Survey was active from 13 July 2020 till 9 November 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Survey for observers' responses

Response rate: 96%



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Survey for Observers

- Improved version of the survey for lexicographic partners
- 121 questions in 6 sections:
  - (1) General information
  - (2) Types of lexicographic resources. Software and tools supporting the workflow
  - (3) Publication and access. Crowdsourcing and Gamification
  - (4) Retrodigitised dictionaries
  - (5) Data formats. Metadata. Availability
  - (6) Past and Future.
- "yes/no" questions, multiple choice questions, open-ended questions. Not all questions were obligatory.
- One survey per institution
- Survey was active from 13 July 2020 till 9 November 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

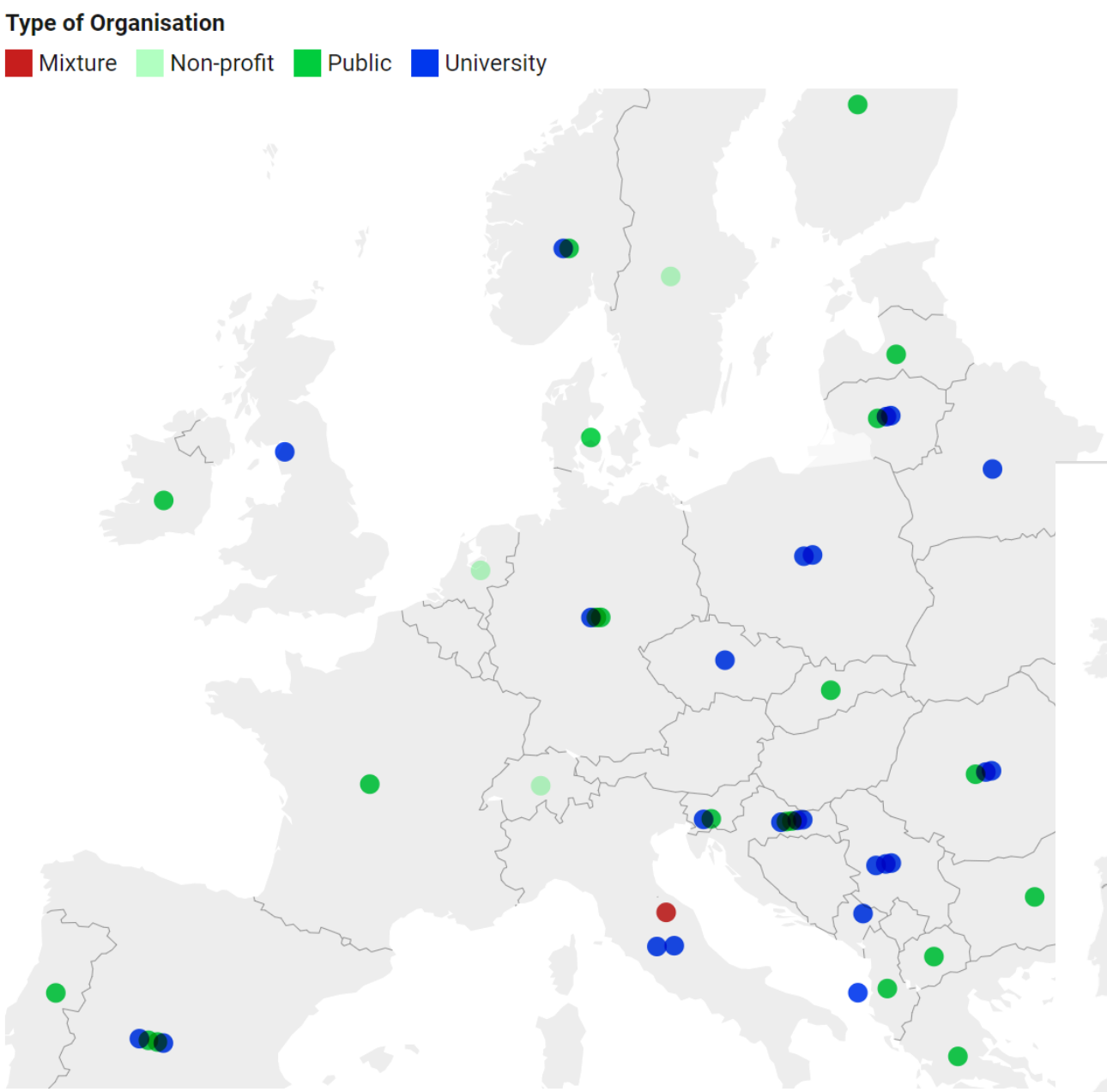
Results - General information

- Once a lexicographer, always a lexicographer
  Most respondents have a PhD in language, and linguistics, and work for at least 6 years in lexicography
  (partners > 20)

- However, most lexicographers do not work exclusively on lexicographic projects
  Even more than 50% of their time spent on other tasks

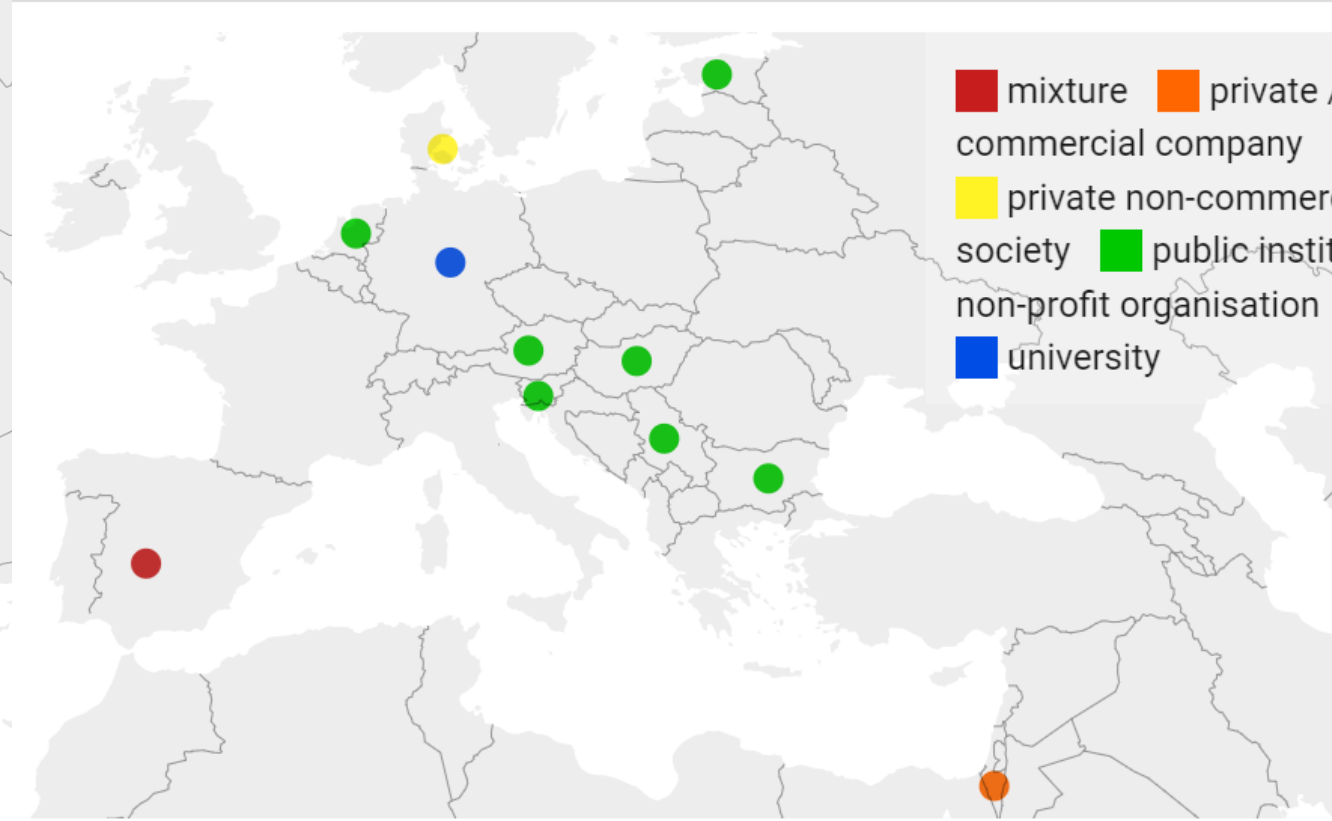
- If trained, most lexicographers are trained within their institution

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



Observer Institutions

Partner Institutions

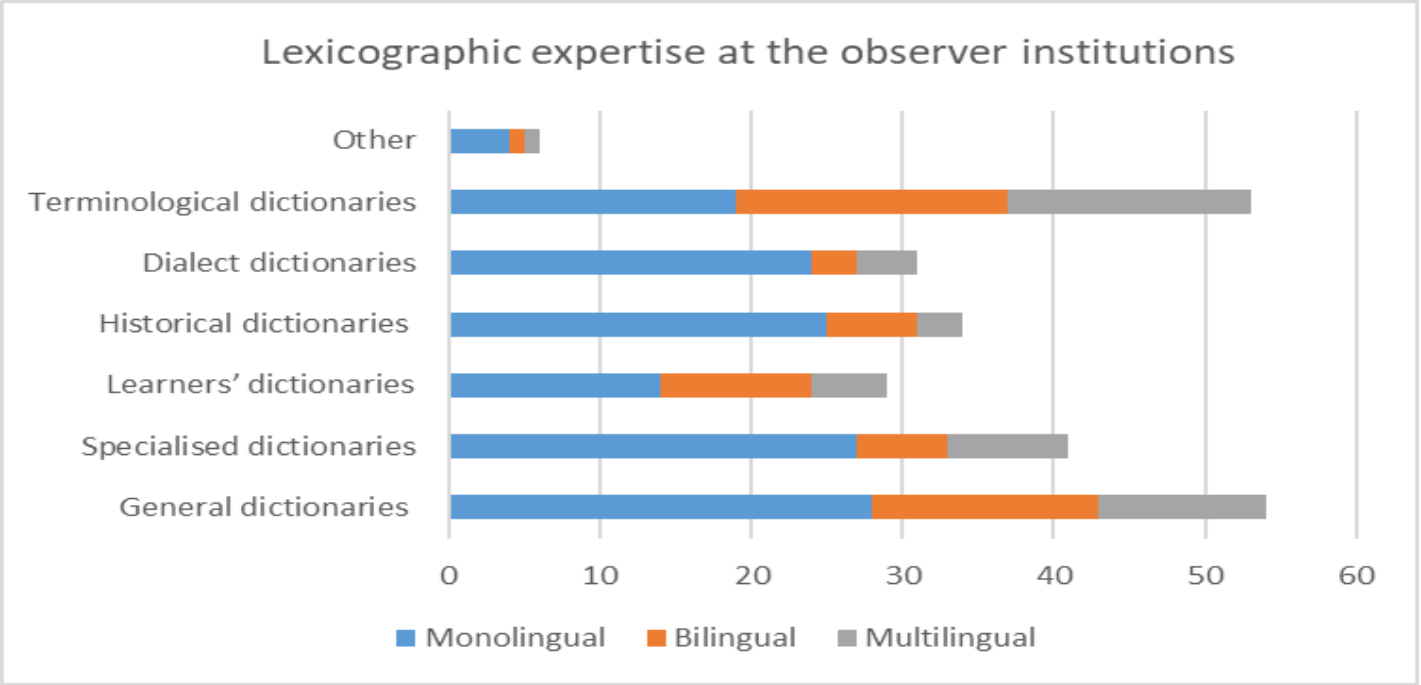


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results – General Information

- Observers have varied lexicographic expertise

- Terminological dictionaries more represented among observers than among ELEXIS partners



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results – General information

- Lexicographic projects are heavily dependent on national funding by the government
  - Monolingual lexicographic work in Europe is mainly done in public institutions (see e.g. Kallas et al. 2019; Kosem et al. 2018)
  - Bilingual, multilingual lexicography more at universities

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results - Software and tools

- Dictionary Writing Systems
- Corpus Query Systems

A lot of different systems: commercial, open-source, in-house

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Corpus Query Systems

A Corpus Query System is a piece of software that lets you see concordances for any word, phrase or grammatical construction in a text corpus, an extensive collection of texts, usually annotated with additional information about words such as their base form or lemma, part-of-speech category and similar.

36 institutions use a CQS, whereas 16 do not (7 of those feel that they need one urgently)

Overall, institutions are satisfied with the CQS

Additional wishes/functionalities:

- advanced corpus creation and annotation tools
- better metadata management
- novel functionalities, e.g. sense clustering, sense annotation and disambiguation, diachronic analysis, detection of translation equivalents
- improved user ergonomy and customisation of the user interface according to user profile, e.g. CQS for learners

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary Writing Systems

a piece of software for writing and producing a dictionary. It might include an editor, a database, a Web interface and various management tools - for allocating work etc. Specialised dictionary editing software includes customisations of existing/standard (XML) editors

21 institutions use a DWS, whereas 31 do not (14 of those feel that they need one urgently)

Institutions are more or less satisfied with the DWS, but some have concerns about the long-term sustainability of the system and keeping up with technical improvements

Reasons for not using a DWS

- financial difficulties
- absence of knowledge and technical skills

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Dictionary Writing Systems continued

Features that are appreciated:

- availability of support
- customisation options (schemas, DTDs and menus, search options, export options)
- the possibility to adapt and add functionalities
- the ability to work with multiple users
- a real-time updating of the database

Additional wishes:

- easy installation
- support for interlinking lexical entries
- adding multimedia files
- API access
- providing links to corpus examples and metadata
- need for publishing policies and licensing regulations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Note on Software and tools

Quite a few observers mention:

- a lack of information regarding usability and effectiveness of available commercial and open-source CQSs and DWSs
- the availability of documentation and training materials as preliminary requirements for adopting a particular solution

See: ELEXIS curriculum on Dariah-Campus

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results - Retrodigitisation

Retrodigitisation is more often practised in specialised lexicographic centres than at universities

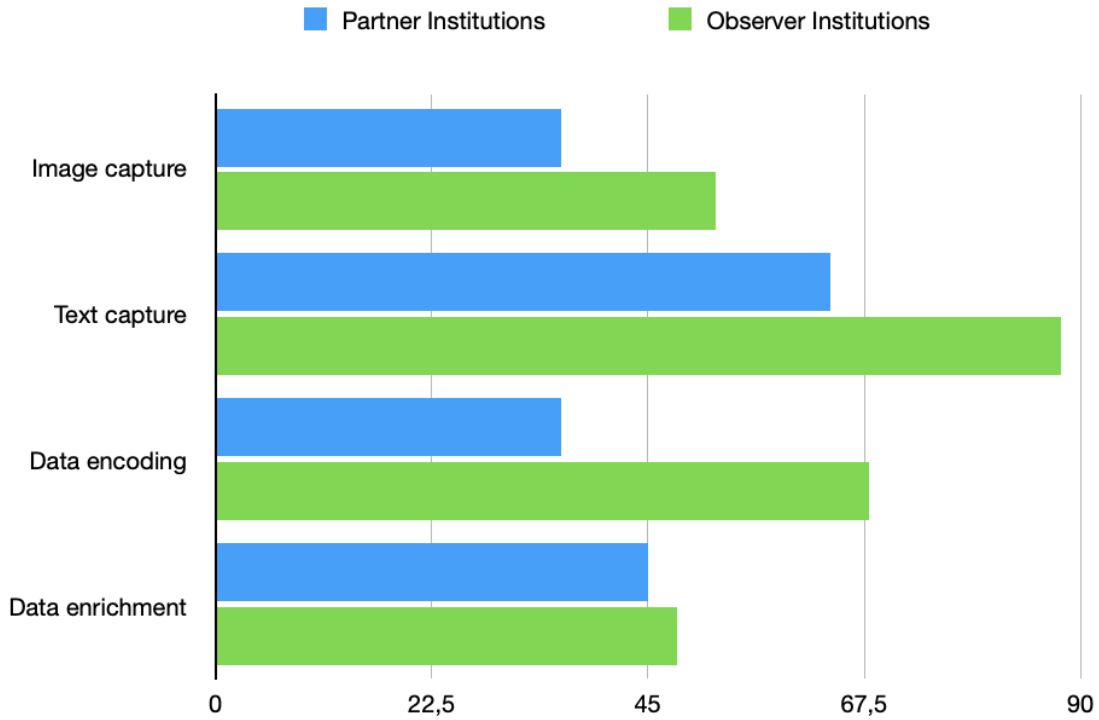


Image capture (using scanners or cameras)
Text capture (OCR or keying (i.e. typing), proofreading, etc.)
Data encoding (structural, i.e. semantic markup, using (TEI) XML)
Data enrichment (e.g. value normalisation, geo-locating, content expansion)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Note on Retrodigitisation

- Interest observed in the whole lexicographic community
- Similar procedures and tools are mentioned for the different phases of retrodigitisation by observers and partners
- Suggests that some best practices are already in place for the retrodigitisation workflow

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results - General Observations

Lexicographic institutions become more of a data provider and less of a dictionary publisher

Modern lexicographer does much more than editing dictionary entries

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results – Wishes for the future

- increased interoperability, linking and sharing of resources
- aggregating stand-alone lexicographic (and also terminological) resources into dictionary portals
  = *Dictionary Matrix*
- (**more**) stable and established formats for data encoding in lexicographic projects = *OASIS LEXIDMA*
- **more** open-source programs and platforms as well as training on how to use them = *Lexonomy & Elexifier & ELEXIS Curriculum*
- **more** NLP resources for low-resourced languages
- intensive integration of lexicographic data into the Semantic Web, AI, and NLP applications

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Results – Concerns and obstacles

- funding
- the low status of lexicographic work
- the low quality and reliability of (semi-)automatically built resources while high quality lexicographic data is still kept under restrictive licenses
- restricted access to the source data for reuse by others

Although serious efforts have been made within ELEXIS to address licensing issues, more promotion and raising awareness seems to be needed to open up lexicographic data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

ELEXIS association

- ELEXIS project ends on 31st July 2022 (no further direct funding)
- Current proposal: ELEXIS Association
  - The objective of this Association is organisation and coordination of activities related to lexicography, and activities related to natural language processing tasks on the topic of semantics, insofar they are of interest to lexicography.
  - Interesting particularly for those who would like to:
    - be actively involved in further development of (lexicographic) tools and services, and exchange of (lexicographic) data (therefore not only providing knowledge about lexicography, as in the case of ELEXIS-KC)
    - search for funding options
    - general collaboration
  - JSI prepared a light-weight „partnership agreement“

    Contact: Simon Krek

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Thank you for your attention

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

Kallas, J./Koeva, S./Kosem, I./Langemets, M./Tiberius, C. (2019a): ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs. https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (last access: 09-07-2022).

Kallas, J./ Koeva, S./Langemets, M./Tiberius, C./Kosem, I. (2019b): Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs. https://doi.org/10.5281/zenodo.3726841 (last access: 09-07-2022).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.